Expanding the Biology Toolkit

Fun with Chemistry

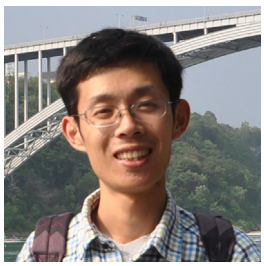
Carolyn Bertozzi
Stanford University & HHMI

Whether you call it biochemistry, molecular pharmacology, or chemical biology, one thing we can agree on is that chemists have long sought to advance the biological sciences. Through development of reagents, instruments, algorithms, and technologies, chemistry brings to biology the ability to drill down to molecules, bonds and atoms—the scale of matter at which all living things converge on common principles. But one often hears the dogma that "you can teach a chemist to do biology but biologists cannot learn to do chemistry."

Not true. There have been developments at the forefront of chemistry that make chemical technologies eminently accessible to biological researchers. Did you know, for example, that you can perform chemical reactions inside cells or model organisms and transform matter like a card-carrying synthetic chemist? You might use such "bioorthogonal chemistries" to monitor de novo DNA biosynthesis using the reagent 5-ethynyl-2'-deoxyuridine (EdU), protein synthesis with azidohomoalanine (AHA), or glycan synthesis with *N*-azidoacetylmannosamine (ManNAz). You can genetically outfit your protein of interest for selective chemical reaction with small-molecule fluorophores using HaloTag, SNAP-Tag, LAP-Tag, and related chemical innovations. Chemists have even made it possible for biologists to synthesize large proteins by assembly of peptide fragments.

These chemical innovations have been honed for transition into the hands of biologists, sometimes via commercial kits. So open your mind to chemistry—you can, and sometimes should, do it.

Mathematical Laws of Randomness

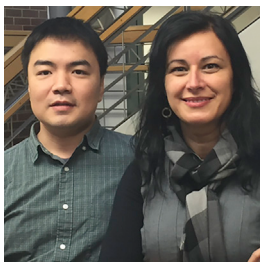


Hao Ge
Peking University

The stochastic processes of transcription and translation inside cells can be described mathematically by a Chemical Master Equation (CME) model, typically simulated by the Gillespie algorithm. Recently, a simple two-state CME model combined with the in vitro single-molecule experiments has revealed the molecular basis for the transcriptional burst under an induced condition in prokaryotic cells (Chong, S., et al., *Cell* 158, 314–326).

The large deviation principle, a highly sophisticated mathematics theory developed only in the late last century, provides an energy-like function (called landscape function) characterizing the non-equilibrium dynamics of living cells. The landscape function provides a rate formula for the phenotype transition—very similar to Arrhenius equation for describing the chemical reaction taught in most college chemistry classes. Inspired by many recent experiments, this general framework has recently been applied to the case in which the gene-state switching is neither extremely slow nor exceedingly rapid (Ge, H., et al., *Phys. Rev. Lett.* 114, 078101). This rate formula nicely explains a "transcriptional noise enhancer" therapy for HIV reactivation: why a class of small molecules that enhance the gene-expression fluctuations but keep the mean transcriptional activity unaltered can significantly reactivate the latent cells (Dar, R.D., et al., *Science* 344, 1392–1396). Finally, the most probable transition path between phenotypic states in multi-dimensional stochastic models, usually not possible to be accurately estimated intuitively, can be numerically obtained, illustrating the power of mathematical methods in understanding random biological processes.

Computing Power for Genomics



Jian Ma and Olgica Milenkovic
University of Illinois at Urbana-Champaign

Genomics research is undergoing a paradigm shift thanks to the development of a myriad of new high-throughput systems for massive data acquisition. These data come with the promise of unprecedented insights into fundamental molecular and cellular mechanisms and the potential for developing models that explain how genomes and regulatory networks function during development and how they differ across species and change in disease state.

Unfortunately, the intrinsic value of such multimodality data is usually unknown, as the systems studied are highly complex, dynamic, and stochastic. Hence, the following question emerges: what methods should we use to evaluate the statistical sufficiency of the data and make the most informative and accurate inference and prediction? To address this question, computational scientists have to engage with biologists in a dialogue, which can be jump-started by a number of exciting ideas in "evidence-based" statistics, information theory, machine learning, and computer science. Small sample detection/estimation, information theory, and graphical models may enable us to understand fundamental inference limits; (causal) compressive sensing matrix and tensor methods may enable the use of sparsity priors; correlation clustering may help in identifying key biological network modules while rank aggregation and prioritization may help with both removing varying data scales and designing biological experiments; deep learning algorithms may enable unprecedented model development.